A NOVEL APPROACH TO CHARACTERIZING CELL MIGRATION

by
Allan B. Yue

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
May 2019

© 2019 Allan B. Yue
All rights reserved

Abstract

Cancer cell metastasis has been responsible for the vast majority of cancer-related deaths in the United States. The processes involved in cancer cell metastasis, such as extravasation and intravasation, are driven by cell motility. The conventional method of characterizing cell motility typically involved imaging live cells under a microscope for several hours and tracking cells' trajectories manually with the help of computer software, and this method is highly inefficient and time-consuming for obtaining cell motility information. This dissertation aims to develop a new method to quantitatively characterize cell motility based on the spatial distribution of cells in clones at a specific time point. A simulation study was first performed to evaluate the correlation between cell spatial distribution in the clones and cell motility. Clonal distributions of cells were generated at the 72 hr time point from computer-simulated cell trajectories based on the PRW model for clone sizes of 4, 16, 32, 64, and 128 cells and effective diffusivities. Then, the spatial distribution of cells were characterized using several parameters including the mean squared displacement from the center of the clone, distances to the center of the clone, and minimum pairwise distances of each cell. The correlation between these parameters and diffusivity was used to identify the parameters that best predict cell motility, and the results showed that the mean distances to the center of the clone best predicted cell motility, with a Pearson's correlation coefficient of 0.969 and an average accuracy of 37.56% across all clone sizes. This accuracy was comparable to that from the PRW model, which had an accuracy of 40.15% over 8 hrs and 28.70% over 16 hrs. Finally, this model was used to analyze an image of MDA-MB-231 cells, which had a diffusivity value of 2.65 ± 1.43 $\mu m^2/min$. This was in range with the experimentally obtained diffusivity values ranging from 0.35 to 9.49 $\mu m^2/min$. The data

showed that analyzing clonal distributions of cells can be effective in characterizing cell motility without needing to manually track cells.

Readers: Professor Denis Wirtz and Professor Pei-Hsun Wu

Acknowledgements

Professor Pei-Hsun Wu and Professor Wirtz have been the best advisers imaginable. Not only did they help supervise my research efforts, but they also implanted in me a sense of curiosity and fearlessness in approaching difficult problems and have taught me a lot throughout the process.

I would also like to thank Ashley Kiemen, a current PhD candidate in the Wirtz lab, for her patience, time, and effort in helping me and answering all of my questions. I would also like to thank her for her efforts in writing the code to help me answer some of my most complicated questions as I was doing the experiments in the lab.

I would also like to thank my parents for supporting me and encouraging me in every step of the way. Without their help and guidance, I would not be able to accomplish what I have done today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Historically, cancer has been among the leading causes of death in the United States. It is largely attributed to the metastasis of cancer cells, which refers to the migration of cancer cells from a primary tumor site to a secondary tumor site. This is a complex, multistep process that is responsible for the vast majority of cancer-related deaths [1]. A schematic of the process is shown below.

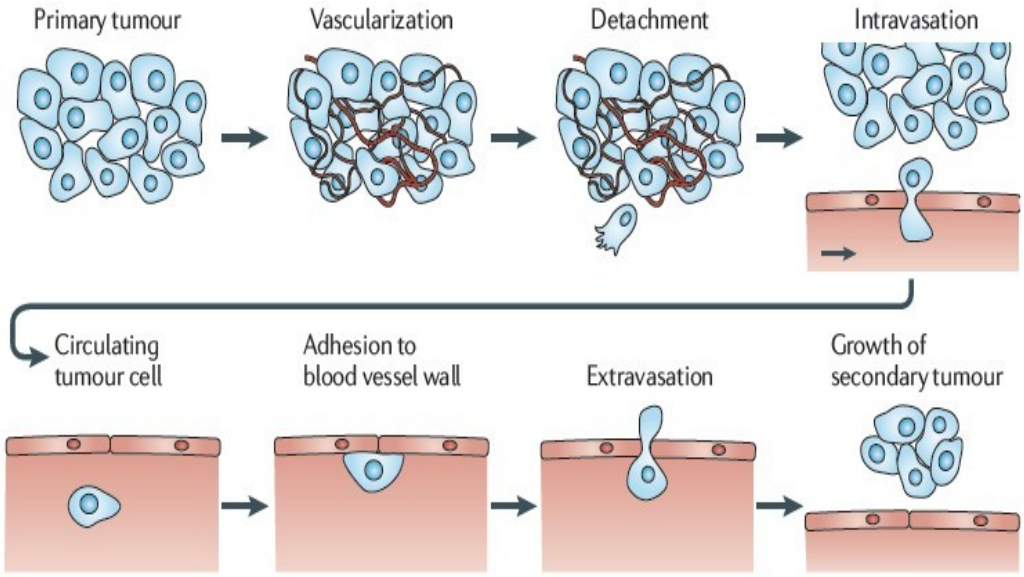

Figure 1.1. A schematic of the cancer cell metastasis process [1].

The cancer cell will vascularize and form blood vessels in the primary tumor site, and then the cell will detach from the primary tumor site, penetrate through the surrounding tissues and blood vessels via intravasation, and circulate inside the blood vessels. Then, the cancer cells will

undergo adhesion to the blood vessel wall, penetrate out of the blood vessels through the extravasation process and into the local tissue, and form a tumor at the secondary tumor site. In both the extravasation and intravasation process, the cancer cells will undergo deformations before penetrating the blood vessels [1]. The processes involved in the metastasis of the cancer cells, such as detachment from the primary tumor cells, intravasation and extravasation, are driven by cell migration [2].

This work explores a novel approach towards characterizing cell motility. So far, existing methods to characterize cell motility are inefficient and time-consuming, and this paper aims to show a potentially quicker and more efficient approach to characterizing cell motility with accuracy similar to that obtained from conventional methods.

This document is organized as follows. The first chapter (Ch. 1) will focus on the current status of the existing methods used to characterize cell motility and the key theoretical ideas necessary to understand the work. Three chapters (Ch. 2, 3, 4) will focus on the building of the model used to characterize two-dimensional cell migration, the comparison of the results from the model with results obtained from conventional experimental methods, and the ability of the model to quantitatively characterize cell motility of an experimentally obtained image of cells. The last chapter will serve to conclude the dissertation. The appendix will include additional information that would help in understanding the work.

1.1 Rationale for using diffusivity

In this dissertation, diffusivity was used to quantitatively characterize cell motility. As it pertains to cancer cells, diffusivity is a quantitative measurement of the rate in which cancer cells spread in the human body. Typically, cells migrate in a random pattern, and two-dimensional cell motility has been characterized by a persistent random walk (PRW) model [3]. Thus, cell migration is a stochastic process [2], and therefore diffusivity was chosen to quantitatively characterize cell motility.

1.2 Introduction to the PRW model

The PRW model describes a cell's trajectory as a succession of uncorrelated movements [2]. In this model, two-dimensional cell motility is characterized with two parameters: cell speed and persistence time [3]. Persistence time refers to the time between changes in the direction of a cell's movement [4]. This characterization comes from the fitting of the cells' mean squared displacements with the PRW model [3].

Using the cell's speed (in $\mu m^2/min$) and persistence time (in min), the diffusivity value is obtained from the PRW model using the following equation.

$$D = \frac{1}{2} \times P\ (\text{min}) \times \text{S}\ \left(\frac{\mu\text{m}}{\text{min}}\right)^2 \qquad (1) \quad [2]$$

In Equation (1), as shown above, D refers to the diffusivity value from the PRW model, P refers to the persistence time (in min), and S refers to the cell's speed (in $\mu m^2/min$).

1.3 Current status

Previously, the experimental method used to characterize two-dimensional cell motility of cells involved imaging live cells under a microscope for several hours. During the imaging process, images of cells were taken every two minutes for several hours at a time [2]. An example of images of cells obtained from the live cell imaging process is shown below.

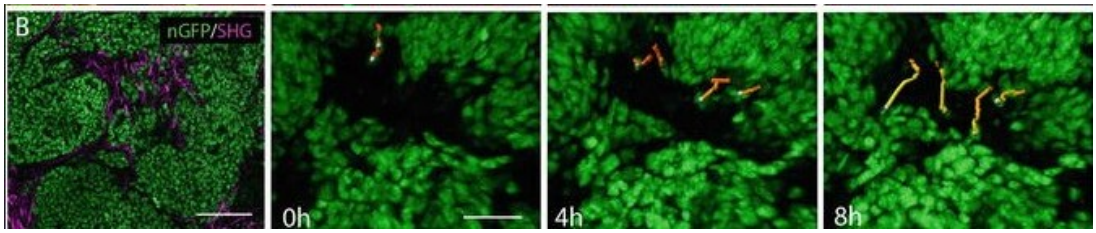

Figure 1.2. Images of intestinal primary tumor cells taken at various time points (0hrs, 4hrs, 8hrs) for 8 hrs [5].

As shown in the above images, the intestinal primary tumor cells were imaged under a microscope for several hours at a time. In this case, the imaging process took 8 hours, and images of the cells were taken periodically during this time.

After the imaging of live cells was complete, the tracking of cells was usually done with the assistance of computer software. Past experiments have used Metamorph to help track the cells' trajectories following the live cell imaging process [2].

The main drawback to using the live cell imaging microscopy method to characterize cell motility is that the process of imaging the live cells under a microscope is time-consuming. Generally, the imaging process takes several hours, and in past experiments, the cells were imaged for 8-16 hours, with images taken every two minutes [2]. This paper aims to examine the effectiveness of our new approach towards quantitatively characterizing cell motility compared

to conventional methods. The next chapter (Ch. 2) will focus on the building of a new model and determining the best model to use for characterizing cell motility.

Chapter 2

New Method to Characterize Cell Motility

In this chapter, a new method to characterize two-dimensional cell motility will be proposed. This method required only one snapshot of cells at a fixed time point to quantitatively infer the diffusivity values and characterize cell motility. It was hypothesized that using one snapshot of cells at a fixed time point will be a more efficient and quicker way to characterize cell motility compared to conventional imaging analysis methods. The following sections will detail the experimental methods that will be utilized to characterize cell motility via the new approach.

2.1 Obtaining snapshots of cell clonal distributions

The first part involved obtaining snapshots of cell clonal distributions. In this part, 100 computer-simulated PRW cell trajectories were obtained for various persistence times and speeds, with a time step of 2 minutes for each of the simulated PRW cell trajectories. 128 repeated simulations were used for each individual PRW cell trajectory [3]. The following table summarizes the persistence time (min), cell speeds (μm^2/min), and clone size (# of cells in a clone) used.

Persistence Time (min)	**Speed (μm^2/min)**	**Clone Size (# of cells)**
1, 5, 10, 20, 30, 40	1, $\sqrt{2}$, $\sqrt{3}$, 2, 5, 10	4, 16, 32, 64, 128

Table 2.1. Persistence times, cell speeds, and clone sizes used in this experiment.

Afterwards, computer-simulated cell division trajectories were obtained under varying cell division times. An example of a computer-simulated cell division trajectory is shown below.

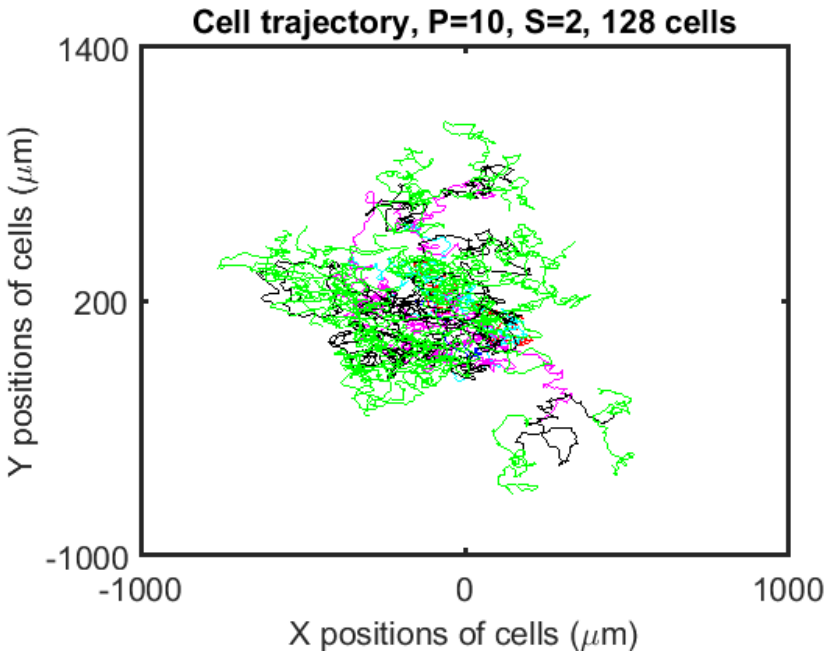

Figure 2.1. Computer-simulated cell trajectory obtained at persistence time 10 mins, cell speed 2 μm^2/min for 128 cells.

As shown above, this cell division trajectory example was obtained for 128 cells. The computer-simulated cell division trajectories were obtained for varying clone sizes, as shown in Table 2.1. The cell trajectories were obtained under different persistence time, cell speed, and clone size conditions in order to assess the effects of cell motility from each parameter.

From the cell trajectories obtained, snapshots of cell clonal distributions at a fixed time point were obtained. An example of a snapshot of cell clonal distributions at a fixed time point is shown.

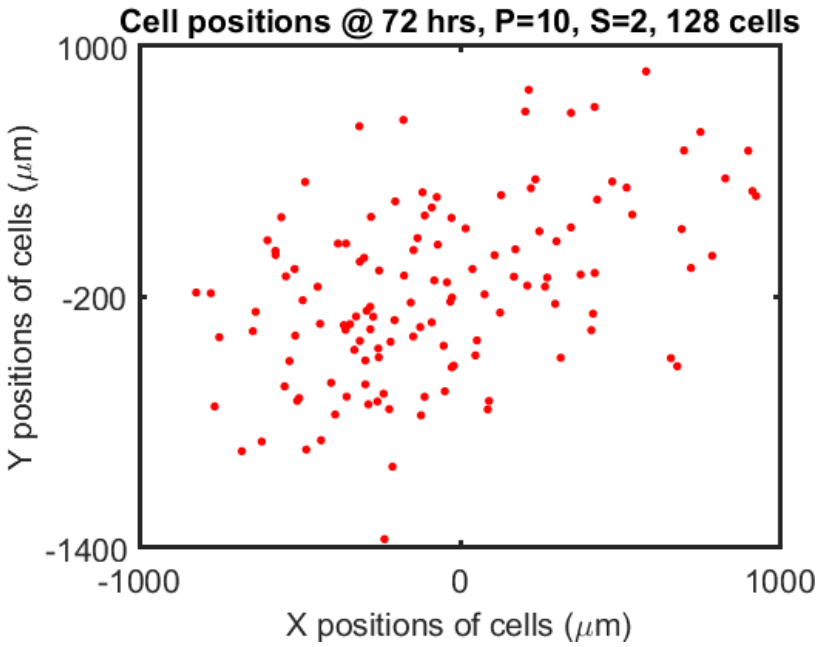

Figure 2.2. A snapshot of the cell clonal distributions obtained at 72 hrs time point, with a persistence time of 10 mins and cell speed of 2 μm^2/min for 128 cells.

These snapshots were obtained at the 72 hours time point. Typically, the human cell division process takes 24 hours [6], however, cell division times vary between different cell types. Therefore, multiple cell divisions were needed to obtain snapshots of clonal distributions with varying clone sizes. It was hypothesized that using snapshots of cell clonal distributions at a fixed time point would be a quicker and more efficient method to characterize two-dimensional cell motility.

2.2 Parameters used to determine best method of characterizing cell motility

In this section, the snapshots of cell clonal distributions obtained at the 72 hours time point were quantitatively analyzed using various parameters, and a correlation analysis of these parameters with diffusivity values obtained from the PRW model in Equation (1) was performed. The purpose of this analysis was to determine the parameters that best characterized cell motility given a single snapshot obtained at the 72 hours time point. The following subsections will detail the steps used to determine the best parameter for characterizing cell motility.

Mean square displacement and distances to the center of the clone

Two parameters that were used to quantitatively analyze the snapshots of cell clonal distributions obtained at the 72 hours time point were mean square displacement from the center of the clone and distances to the center of the clone. To locate the center of the clone of cells, the minimum and maximum values for the x and y coordinates of cells in each snapshot were first obtained. Then, the midpoint of the minimum and maximum values for the x and y coordinates was found. An example of this is shown in the following snapshot.

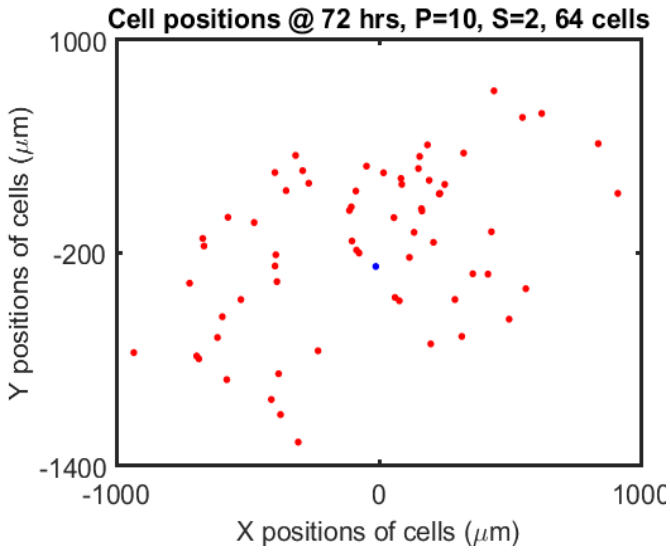

Figure 2.3. A snapshot of the cell clonal distributions at 72 hours time point with center of the clone.

As shown in the snapshot above, the steps used to find the center of the clone of cells were used. The blue dot represents the center of the clone, while the red dots in the snapshot represent the cells in the clonal distribution. Once the clone center was found, the mean square displacement from the center of the clone (MSD) and the distances to the center of clone (DCC) were calculated using the following equations.

$$DCC = \sqrt{(x_n - x_{center})^2 + (y_n - y_{center})^2} \qquad (2)$$

$$MSD = \frac{1}{\#\ cells} \sum_{n=1}^{\#\ cells} (x_n - x_{center})^2 + (y_n - y_{center})^2 \qquad (3)$$

After obtaining the mean square displacements and the distances to the center of the clone, the mean, median, standard deviation, interquartile range, and range of the distances to the center of the clone were obtained. These values were used for the correlation analysis with diffusivity values obtained from the PRW model in Equation (1).

Pairwise distances

Another parameter that was used to quantitatively analyze the snapshots obtained at the 72 hours time point was the pairwise distance. Pairwise distances refer to the distances between a particular cell and its neighboring cells [7], as shown in the figure below.

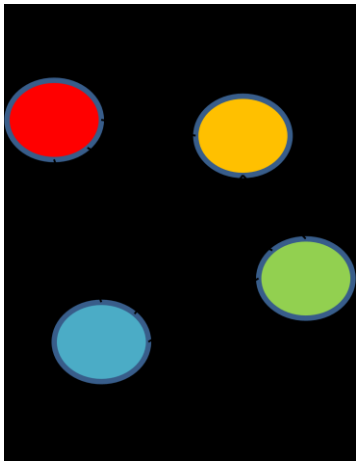

Figure 2.4. An illustration of pairwise distances for four cells.

The figure above shows an example of how pairwise distances were obtained. In this case, pairwise distances were obtained between the red cell (cell 1) and the orange, green, and blue cells (cell 2, 4, and 3, respectively). This procedure was repeated for each cell in the clone. Then, for each cell in the clone, the minimum value of the pairwise distances was obtained using Matlab, and the mean, median, mode and range of minimum pairwise distances were obtained for each clone. These values, along with the mean, median, standard deviation, range and interquartile range of the distances to the center of the clone and the mean square displacement from the center of the clone, were used for the correlation analysis with diffusivity values obtained from the PRW model using Equation (1).

2.3 Correlation analysis

The next step in determining the best method to quantitatively characterize cell motility from a snapshot at a fixed time point was to perform a correlation analysis of all of the parameters obtained. Using Matlab, these parameters were analyzed to see how well each of them correlated with the diffusivity values from the PRW model. The following table summarizes the results of the correlation analysis.

Measured Observation	Statistical Parameter	r Value
Minimum pairwise distance	Mean	0.855
	Range	0.769
	Median	0.779
	Mode	0.507
DCC	Mean	0.969
	Median	0.969
	Interquartile Range	0.902
	Range	0.841
	Standard Deviation	0.911
MSD		0.968

Table 2.2. Results of the correlation analysis of various parameters with diffusivity values obtained from the PRW model.

Based on the results of the correlation analysis shown in Table 2.2, it was determined that the mean and median distances to the center of the clone, along with the mean square displacement from the center of the clone, correlated best with the diffusivity values. Thus, it is predicted that these parameters were able to quantitatively characterize cell motility well. The plots of the correlation of these parameters with the diffusivity values are shown below.

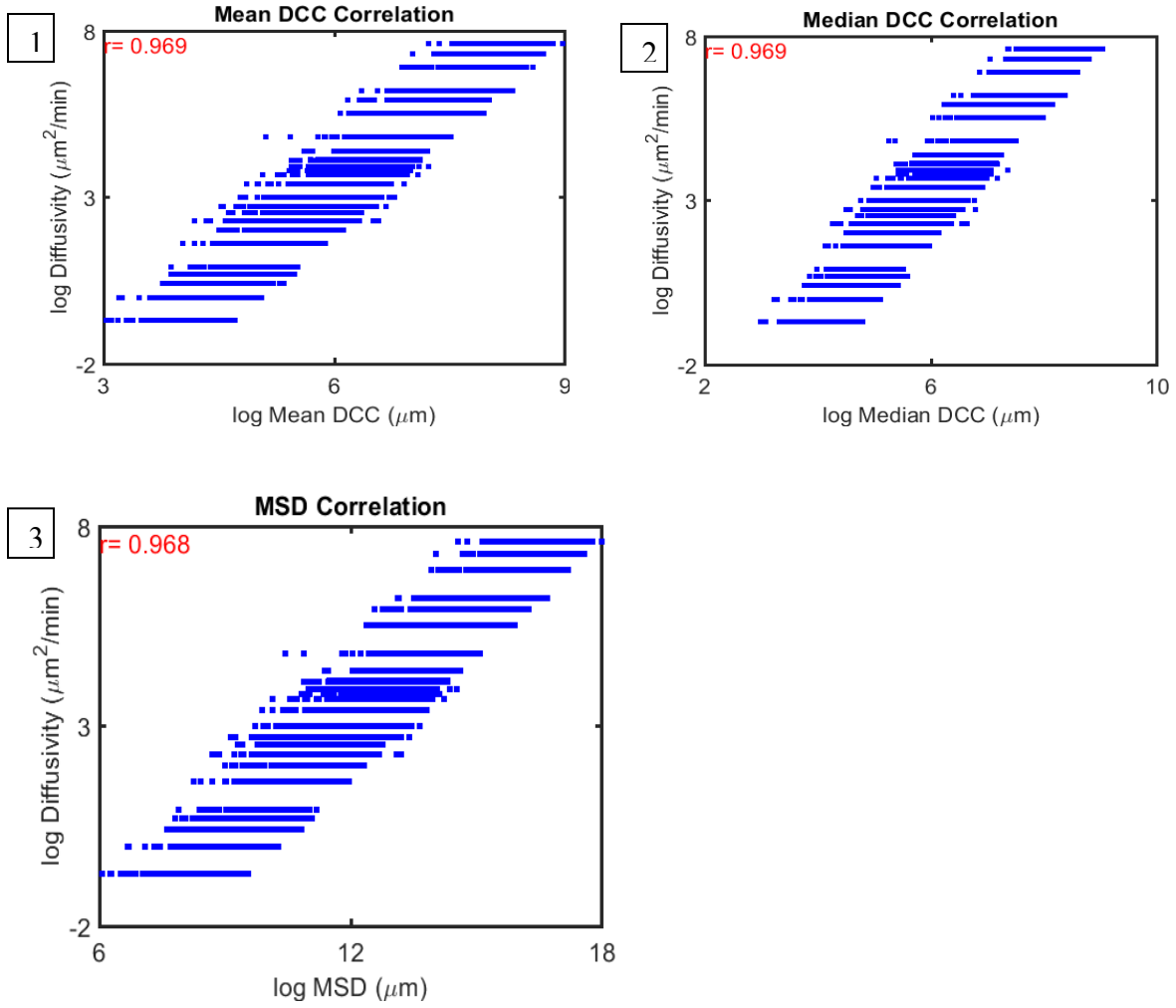

Figure 2.5. Plots of mean DCC (1), median DCC (2), and MSD (3) correlation with diffusivity values.

2.4 Calculating diffusivity values

Using the mean DCC, median DCC, and MSD from the center of the clone, the next step was to obtain the models needed to calculate diffusivity values. To do this, these values were fitted with the diffusivity values obtained from the PRW model in Equation (1) using Matlab. The following table shows the linear models used to calculate the diffusivity values from these parameters.

Observation	Linear Model Used
Mean DCC	$\ln(\check{D})=1.88\ln(\text{mean DCC})-8.32$
Median DCC	$\ln(\check{D})=1.89\ln(\text{median DCC})-8.32$
MSD	$\ln(\check{D})=0.94\ln(\text{MSD})-8.44$
MSD (theoretical)	$\ln(\check{D})=\ln(\text{MSD})-9.76$

Table 2.3. Linear models used to calculate diffusivity values from MSD from clone center, mean DCC, and median DCC.

As shown in the table above, the fitting of the parameters was done in log-log in order to address the issue of skewness towards large values [8]. In this experiment, the diffusivity values ranged from as low as 0.5 to as high as 2000 μm^2/min, and this was attributed to the skewness towards large values when the fitting of the parameters was done. For log-log fitting, the natural logarithm values were obtained for each of the parameters, fitted with the natural logarithm of the diffusivity values obtained from the PRW model, and analyzed. The following table compares the R^2 values between log-log fitting and non-log-log fitting.

Observation	Linear Model R^2 Values	
	Logarithm	Non-logarithm
Mean DCC	0.94	0.832
Median DCC	0.94	0.825
MSD	0.937	0.819

Table 2.4. Comparison of R^2 values between log-log fitting and non-log-log fitting.

As shown above, the linear models obtained from log-log fitting had higher R^2 values than the linear models obtained from non-log-log fitting. This indicates a better fit of linear models obtained from log-log fitting.

The linear models obtained from log-log fitting, shown in Table 2.3, were used to calculate diffusivity values. The theoretical linear model used to calculate diffusivity values from the mean square displacement from the center of the clone comes from the natural logarithm version of the formula MSD = 4 × Ď × time. Then, the accuracy of the results was obtained by comparing the diffusivity values obtained from the linear models with the diffusivity values obtained from the PRW model. The following table summarizes the results.

Method of Diffusivity Calculation	% Error of Calculated Diffusivity			
	Mean	Minimum	Maximum	Range
Mean DCC	37.56%	0.01%	479.29%	479.27%
Median DCC	38.36%	0.001%	573.20%	573.20%
MSD	38.43%	0.02%	418.41%	418.39%
MSD (theoretical)	38.36%	0%	211.34%	211.34%

Table 2.5. Accuracy of calculated diffusivity from the linear models.

As shown above, mean DCC had the lowest mean percent error in calculating diffusivity, but it was slightly lower than the other methods. Mean DCC also had a higher range of percent error compared to that from the theoretical linear model used to calculate diffusivity values from the MSD. This is attributed to the fact that cell migration is a stochastic process [2], and that cell movements occur in a random pattern [3].

2.5 Random selection of clones

In this section, the goal is to determine the number of clones needed to calculate diffusivity values and quantitatively characterize cell motility from the snapshots with reasonable accuracy. The table below summarizes the number of clones used in this section.

# Clones Used for Random Picking (N)	10, 20, 30, 50, 100, 200, 300, 400, 500, 600, 1000

Table 2.6. Number of clones used for the random picking analysis.

10 clones were initially chosen for this analysis, and then the mean DCC was obtained for each of these 10 randomly chosen clones. Afterwards, the diffusivity values were obtained using the linear model in Table 2.3. Then, the average values of the diffusivity values obtained from the PRW model in Equation (1) and of the diffusivity values obtained from the linear model in Table 2.3 using mean DCC were taken over the 10 randomly selected clones, and the accuracy of the diffusivity values was obtained. This procedure was repeated over 100 trials, and the average accuracy was obtained over 100 trials. These steps were then repeated for higher numbers of clones, as shown in Table 2.6. Afterwards, the confidence intervals were taken for the mean percent error values obtained over 100 trials for each number of randomly selected clones, at the 95% confidence level. The results of this analysis are summarized in the following table.

# of Randomly Chosen Clones	Avg % Error of Calculated Diffusivity from Mean DCC	
	Mean	Confidence Interval Length
10	19.70	6.27
20	14.90	4.49
30	13.58	3.87
50	11.25	3.95
100	10.92	2.56
200	8.80	2.11
300	8.55	1.66
400	8.42	1.48
500	8.74	1.38
600	8.21	1.33
1000	8.94	0.98

Table 2.7. Mean percent error of calculated diffusivity and 95% confidence interval for each number of randomly selected clones.

Given the above results, 200 clones were necessary to calculate diffusivity values and characterize cell motility with a reasonable accuracy. The plots below illustrate these results.

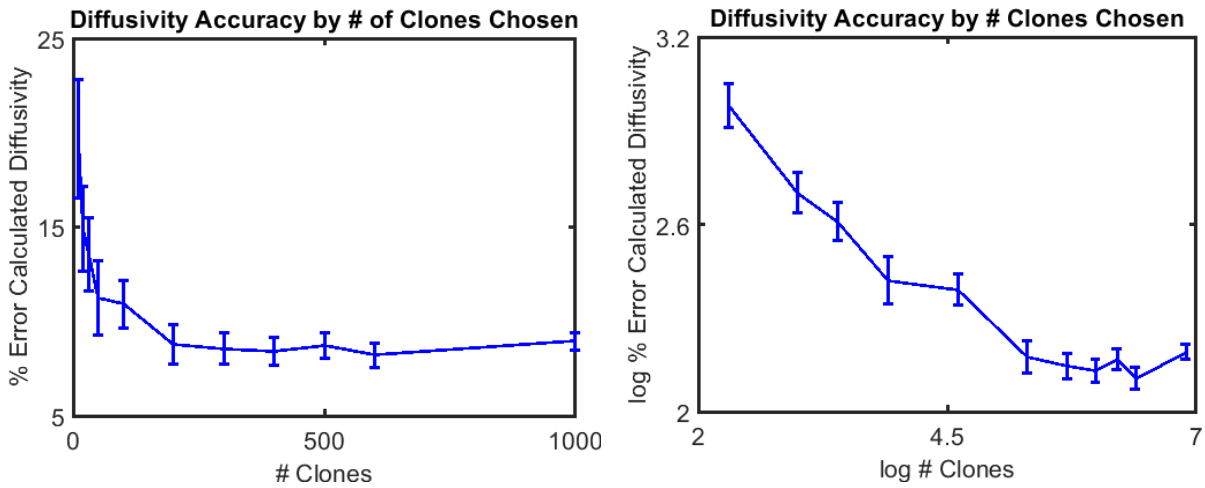

Figure 2.6. Accuracy of diffusivity values from mean DCC by number of randomly chosen clones, in non-log-log (left) and log-log (right).

The accuracy of calculated diffusivity from mean DCC, as shown above, exponentially decreased from 10 clones to 200 clones, and did not change much when more than 200 clones were randomly selected. The next chapter (Ch. 3) will show how the results obtained from the new approach compared to results obtained from conventional approaches.

Chapter 3

Comparing the New Method with Conventional Methods

In this chapter, the results from the new method will be compared to the results from conventional methods to assess whether the new method quantitatively characterized cell motility more quickly and efficiently. The following sections will detail the procedures carried out to compare these results.

3.1 Obtaining PRW cell trajectories

The first step was to obtain the PRW cell trajectories. 100 computer-simulated PRW cell trajectories were obtained for each persistence time and cell speed [3], and the persistence times and cell speeds combinations used are shown in Table 2.1.

This procedure was carried out over 8 hrs and 16 hrs time period, with a time step size of 2 mins, and cell trajectories were obtained from each of these conditions. An example of a PRW cell trajectory obtained over 16 hrs time period is shown.

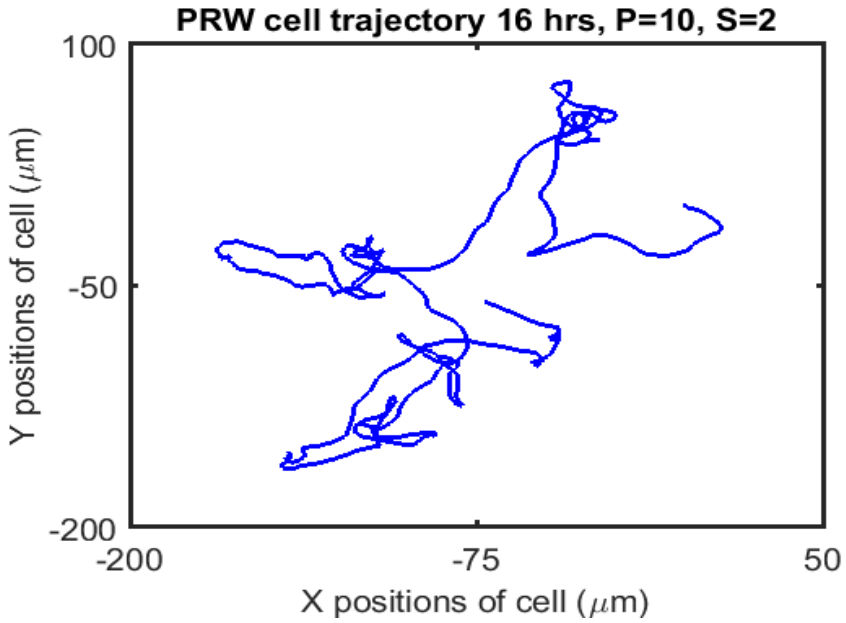

Figure 3.1. PRW cell trajectory obtained over 16 hrs time period, persistence time 10 mins and cell speed 2 μm/min.

3.2 Diffusivity calculations

After the PRW cell trajectories were obtained, MSDs from the cell trajectories were fitted with the PRW model to obtain the fitted persistence times and speeds according to the following equation below.

$$MSD(\tau) = 2S^2P(\tau - P(1 - e^{-\tau/P}) + 4\sigma^2 \quad (3) \quad [2]$$

After obtaining the fitted persistence times and speeds, these values were used to obtain the diffusivity from the PRW model using Equation (1). Then, these diffusivity values were compared with the general diffusivity values obtained from the PRW model using the same persistence times and speeds (see Table 2.1) that were used to obtain computer-simulated PRW cell trajectories. The mean percent error of calculated diffusivity values was obtained, and the

following table summarizes the comparison of the results obtained from mean DCC versus the results obtained from conventional methods (using the PRW model).

Method of Diffusivity Calculation	Mean % Error Calculated Diffusivity
Mean DCC	37.56%
PRW 16 hr	28.70%
PRW 8 hr	40.15%

Table 3.1. Comparison of accuracy of calculated diffusivity between mean DCC and PRW model methods.

As shown above, using the mean DCC to calculate diffusivity and characterize cell motility was quicker and more efficient than the PRW model method. The accuracy of calculated diffusivity obtained from the mean DCC was similar to that obtained from the PRW model methods. The following heat maps provide more detailed analysis of these results.

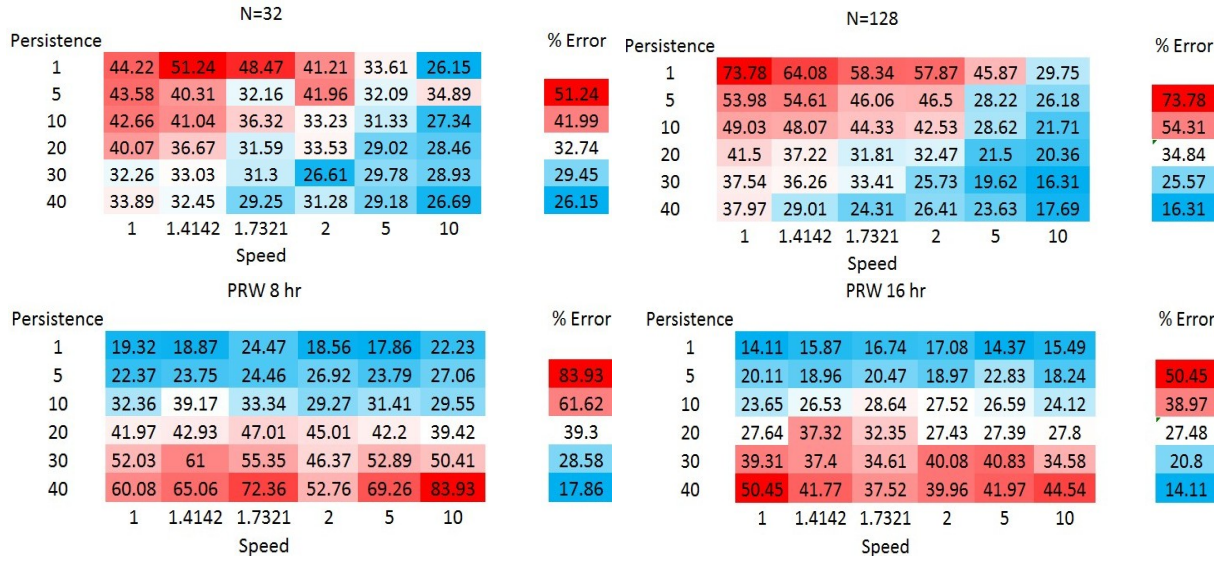

Figure 3.2. Heat maps of accuracy of calculated diffusivity for various persistence times and cell speeds, obtained from clone size of 32 cells and 128 cells (top) and from PRW model at 8 hrs and 16 hrs (bottom).

According to the heat map above, results from the PRW model showed lower percent error of calculated diffusivity values at lower persistence times. On the other hand, results from the method of using mean DCC to calculate diffusivity from a snapshot of a clone of cells at a fixed time point showed lower percent error of calculated diffusivity values at higher persistence times and cell speeds. The next chapter (Ch. 4) will focus on applying the new method to analyzing images of cells obtained experimentally from the lab.

Chapter 4

Using the new method to analyze images of cells

In this chapter, the new method will be used to analyze the images of cells obtained from the lab. The expectation is that the new method could obtain reasonably accurate diffusivity values of the cells from the image more quickly and efficiently compared to that obtained from conventional methods. The following sections will detail the steps used to analyze the images of cells.

4.1 Finding the cell locations from the image

In this analysis, one experimentally obtained image of MDA-MB-231 cells from the lab was used. This image was stained using DAPI (4',6-diamino-2-phenylindole), which is a blue-fluorescent DNA stain that binds to the adenine-thymine region and fluoresces. Although DAPI is typically used as a DNA stain, it can also be used for live cells at a higher concentration [9]. The image obtained from the lab is shown in the following figure.

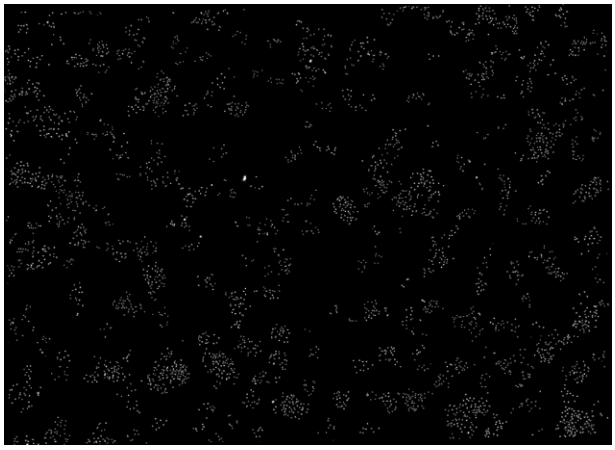

Figure 4.1. DAPI stained image of MDA-MB-231 cells obtained from the Wirtz lab.

Using the image obtained in Figure 4.1, the coordinates of the cells were located using ImageJ-Fiji's Cell Counter feature, where the cells were counted and the locations of the cells were found. Then, the locations of these cells in the image were saved and exported into a Microsoft Office Excel 2007 spreadsheet, in XLS file format. Afterwards, this file was loaded into Matlab, and the image of MDA-MB-231 cells was produced using the x and y coordinates of the cells' locations obtained from the ImageJ-Fiji's Cell Counter feature. This image is shown below.

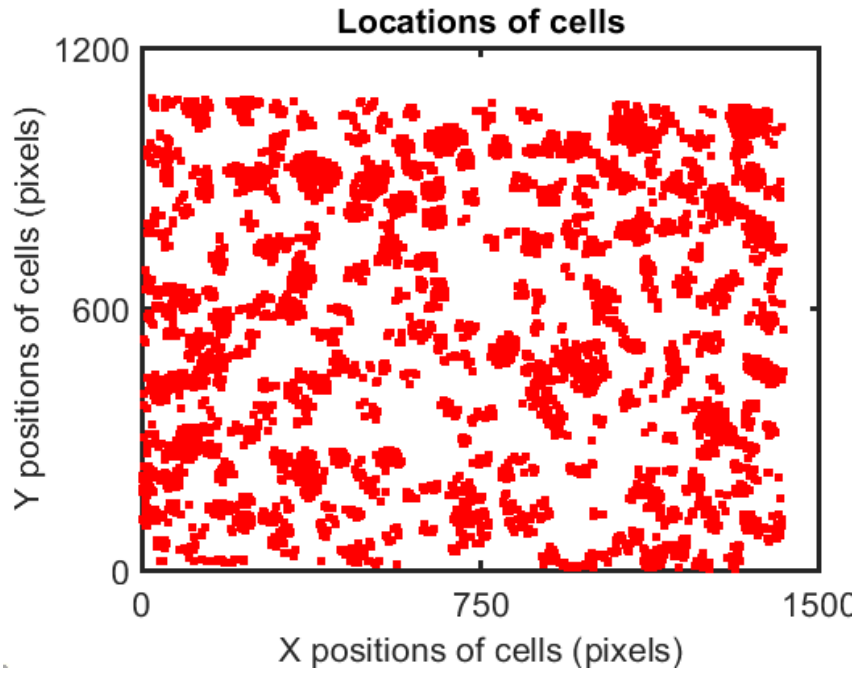

Figure 4.2. Image of MDA-MB-231 cells' locations obtained from the Matlab software.

Using the image in Figure 4.2, the next step was to obtain the clones of cells that were used to quantitatively characterize cell motility from this image. The methods used to obtain the clones of cells will be described in the next section.

4.2 Obtaining clones of cells from the image

In this section, the clones of cells were obtained from the image. The method used to obtain the clones of cells from the image was the Gaussian mixture model, which is a mixture of different Gaussian distribution components. An illustration to show the Gaussian mixture model is shown below.

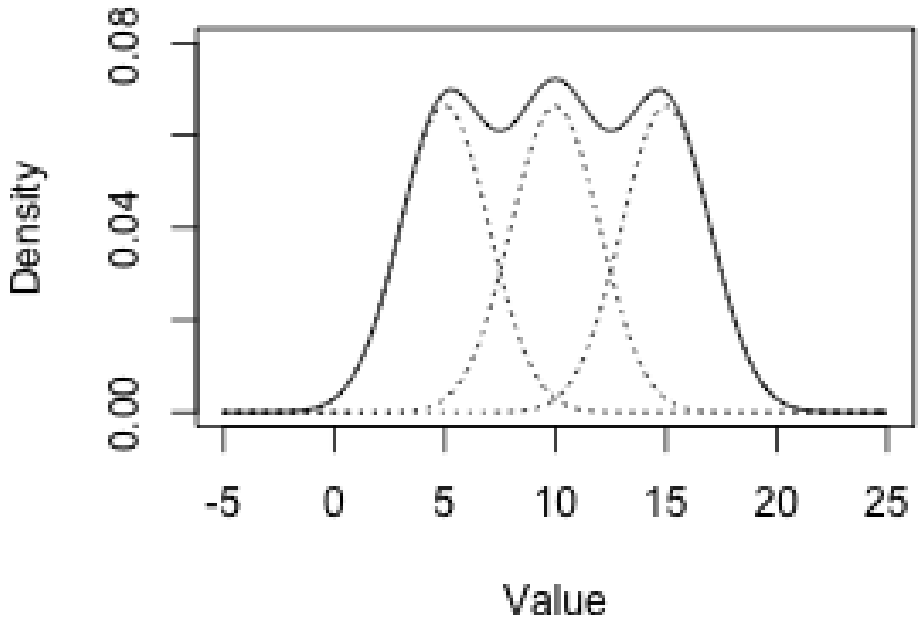

Figure 4.3. Illustration of the Gaussian mixture model [10].

The Gaussian mixture model consists of individual Gaussian distributions, and the mixture model represents a combination of all of the individual Gaussian distribution components. This model was used to obtain clones that were fitted using estimated Gaussian means and random initialization of Gaussians, using Matlab code written by Ashley Kiemen, a PhD student in Professor Wirtz's lab.

4.3 Factors that affected the fitting of clones of cells

There are several factors that affect the fitting of the clones of cells, which affects the grouping of cells into different clones. These factors will be discussed below.

Covariance matrix

The covariance matrix is one important factor that affected how the cells were grouped into different clones. The type of covariance matrix, as well as whether it is shared or not, affects the direction, length, and size of the contours. In a full covariance matrix, the contours adopt a shape independently. However, in a diagonal covariance matrix, the contours are oriented along the axes of the plot [11]. These factors affect the grouping of the cells into different clones, and these differences are illustrated in the plots below.

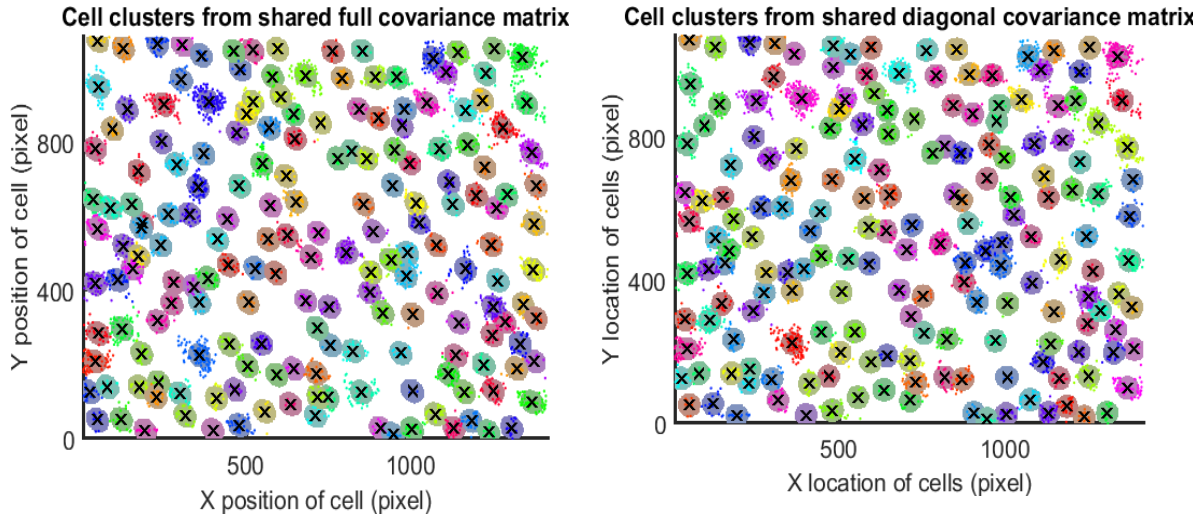

Figure 4.4. Cell clusters obtained using a shared full covariance matrix (left) and a shared diagonal covariance matrix (right).

Additionally, whether or not the components used the same covariance matrix affects the size and orientation of the contours. In general, the contours are of the same shape and size when the components of the Gaussian mixture model share a covariance matrix. On the other hand, the contours have different shapes and sizes when the components do not share a covariance matrix [12]. These factors affect the clustering of the cells into different clones.

Algorithms in the Gaussian mixture model

Another important factor that affects how the clones are fitted in the image has to do with the algorithms used in the Gaussian mixture model. The differences between these algorithms also contribute to differences in the way the cells were grouped into clones in the image. Two methods were used to obtain the fitting of clones in the image of cells. In the estimated Gaussian means method, the locations of the Gaussian mean peaks were located, and the cells' distances from each of the Gaussian mean peaks were calculated. The cells were then grouped into clones based on their proximities to the Gaussian mean peaks.

Another method used to obtain the fitting of clones was the k-means++ method. In this method, a heuristic was used to find the initial seeds that would be used for the k-means clustering method. The k-means clustering method was then used to cluster the cells in the image into a specific number of clones defined by the initial seeds that were found prior to the implementation of k-means clustering. These differences affect the grouping of the cells into clones, as shown in the following plots.

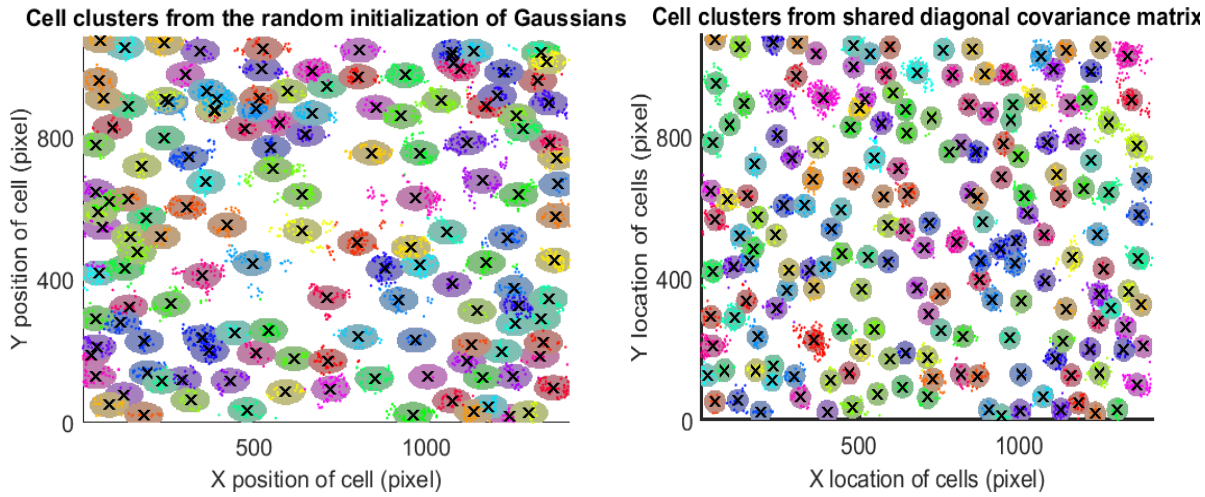

Figure 4.5. Cell clusters obtained using k-means++ method (left) and estimated Gaussian means method (right).

The next step was to find out which plot best fitted with the clones obtained using the Gaussian mixture model. This will be discussed in the next section.

4.4 Finding the best fit plot

In this section, the steps used to analyze the plots and to find the plot that best fitted with the clones obtained will be outlined. Two values were used to find the best fit plot: the Akaike information criterion (AIC) and the Bayes information criterion (BIC). The AIC is based on in-sample fits to estimate the likelihood of a model to predict future values, whereas the BIC measures the trade-off between model fit and model complexity [13]. Ideally, the best fit plot has the lowest AIC and BIC values.

In the earlier sections, three plots of clones fitted with the Gaussian mixture model were obtained. Two of the three plots used the estimated Gaussian means method to fit the clones. In these two plots, one had the shared full covariance matrix while the other had the shared

diagonal covariance matrix. The third plot also had the shared diagonal covariance matrix, however, the algorithm used to obtain the fitted clones was k-means++. These plots were analyzed, and goodness of fit was considered. Using the Matlab software, the AIC and BIC values for each plot were obtained, and the plot with the lowest AIC and BIC values was identified as the best plot that fitted with the clones obtained for the cells in the image. The following table summarizes the results.

Algorithm used	Covariance matrix	AIC value	BIC value	# clusters
Estimated Gaussian means	Shared full	112310	116110	182
Estimated Gaussian means	Shared diagonal	112340	116140	182
k-means++	Shared diagonal	114250	118105	143

Table 4.1. AIC and BIC values of plots of fitted clones obtained from the Gaussian mixture model.

As shown above, the plot obtained using the estimated Gaussian means, with a shared full covariance matrix, had the lowest AIC and BIC values. This indicates that this plot best fitted with the clones obtained from the Gaussian mixture model. Therefore, this plot was used to obtain the diffusivity values and characterize the cell motility in the image of MDA-MB-231 cells. 182 clusters were obtained from the plot as indicated in Table 4.1.

4.5 Calculating the diffusivity values

Using the best fit plot identified in the previous section, the diffusivity values of each clone were calculated. For each clone, the center of the clone was located by obtaining the minimum and maximum values of the x coordinates and of the y coordinates. Then, the midpoint of the minimum and maximum values of the x coordinates and of the y coordinates was found. An example of a cluster obtained from the Gaussian mixture model is shown below.

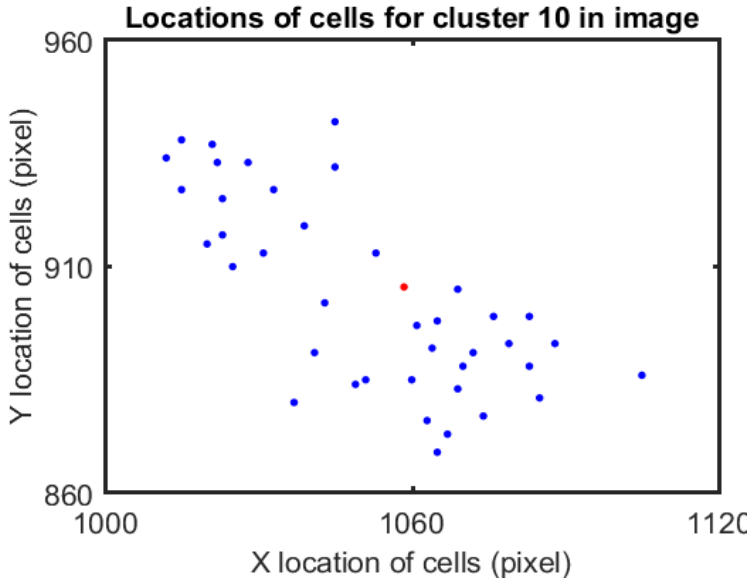

Figure 4.6. Locations of cells for cluster #10 in sample image, with center of clone in red.

Afterwards, the distances of the cells (in blue) to the center of the clone (in red) were calculated, and the mean DCC was obtained for each clone in the image (see Figure 4.6 for an example). Since the mean DCC was obtained in pixels, this value was converted to µm using the 1 pixel = 6.4 µm conversion rate. The diffusivity values for each clone were then obtained from the linear model in Table 2.3 using the mean DCC. Afterwards, the mean diffusivity value was obtained

over 182 clones. The following table compares the results obtained from mean DCC with the results obtained from conventional experiments.

Method of Diffusivity Calculation	Condition	Diffusivity (μm^2/min)
Mean DCC		2.65 ± 1.43
2D PRW model (HT1080)		6.61 ± 5.02
3D PRW model (HT1080)	2 mg/ml	2.02 ± 3.23
	2 mg/ml, Latruculin B, 100 nM	0.39 ± 0.64
	2 mg/ml, blebbistatin, 15 μM	1.00 ± 2.04
2D PRW model (MDA-MB-231)		9.49 ± 13.41
3D PRW model (MDA-MB-231)	2 mg/ml	0.35 ± 0.51

Table 4.2. Comparison of diffusivity results using mean DCC over experimental results [3].

As shown above, the diffusivity values obtained from mean DCC for the image were within range. This indicated that using the mean DCC to characterize cell motility from one snapshot at a fixed time point produced similarly accurate results compared to results obtained from conventional experimental methods more quickly and efficiently.

Chapter 5

Conclusions

This chapter concludes the thesis. The results obtained from the experiments will be summarized and discussed. Lastly, some potential ideas for future studies will be provided.

5.1 On the results

From this experiment, it was concluded that the mean distance to the center of the clone (DCC) had the lowest mean percent error in calculating diffusivity values and characterizing cell motility for a snapshot of a clone of cells obtained at a fixed time point, as shown in Table 2.5. But, the mean percent error of diffusivity values obtained using mean DCC was slightly lower than that obtained using median DCC and MSD. The diffusivity results obtained using mean DCC also had a large range of percent error values. This could be attributed to the fact that cell migration occurred in a random pattern [3], which could lead to differences between the observed diffusivity value and the diffusivity value obtained from the PRW model at the single-cell level.

Mean DCC quantitatively characterized cell motility with reasonable accuracy when compared to that obtained from conventional experimental methods. Past experiments used live cell imaging microscopy techniques. On the other hand, the mean DCC method simply used one snapshot at a fixed time point to calculate diffusivity of cells. This showed that mean DCC can quantitatively characterize cell motility more efficiently and quickly compared to that of conventional experimental methods. Additionally, the mean DCC method calculated diffusivity values that were in range with diffusivity values obtained from past experiments [3]. However,

this did not take into account the biological conditions that took place during the experimental process, which could explain the differences between the diffusivity values obtained from the past experiments and the diffusivity values obtained using the mean DCC method from the image of MDA-MB-231 cells. Diffusivity values obtained using the mean DCC method represent values that would be obtained under ideal conditions. But, the state of the laboratory conditions and of the biological conditions of the cells could impact diffusivity values obtained from the cells in the actual experiments.

Lastly, 200 clones were needed to calculate diffusivity values and characterize cell motility with reasonable accuracy. There was a big difference in the accuracy of calculated diffusivity values between 10 and 200 randomly selected clones. However, when more than 200 clones were randomly chosen, the accuracy of calculated diffusivity values did not change much.

5.2 Future studies

One idea for future studies would include applying the model to images of other types of cells that were experimentally obtained from the lab. This would give a better sense of whether or not the model could obtain similarly accurate results in a quicker and more efficient way.

Additionally, applying the model to images of cells obtained under different biological conditions could potentially provide a better comparison of the diffusivity values obtained from the model over that obtained from conventional experimental methods.

Appendix A

Confidence Intervals

The confidence interval is very useful in determining the range of values that one is certain the true value lied in [14]. The confidence intervals were calculated for the accuracy of calculated diffusivity given a specific number of randomly selected clones, and the following section will outline the specifics of the process.

A.1 Calculation of the confidence interval

In calculating the confidence interval, 100 trials of the accuracy of calculated diffusivity of a given number of randomly selected clones, which was obtained by comparing the mean diffusivity value obtained from the PRW model using Equation (1) with the mean diffusivity value obtained from the mean DCC value using the linear model over a specific number of randomly selected clones from Table 2.3, were performed. Then, the confidence interval was calculated using the following formula.

$$CI = z_{95\%\ confidence}\left(\frac{\sigma_{sample}}{\sqrt{\#\ trials}}\right) \tag{4}$$

The standard deviation of the percent error of calculated diffusivity, obtained from 100 trials for a specific number of clones, was calculated. Then, it was divided by the square root of the number of trials obtained. Afterwards, this value was then multiplied by the z-value for the 95% confidence level. The 95% confidence level is most commonly used in health-related publications [15]. The z-value for the 95% confidence level is 1.96.

To calculate the length of the confidence interval, the error value obtained using Equation (4) was doubled, since the confidence interval spans from values below the true mean to the values above the true mean. These values are shown in Table 2.7.

Following this, the confidence intervals were fitted to the log-log plot. To accomplish this, the confidence intervals were fitted using the following equation.

$$\delta z = 0.434 \frac{\delta y}{y} \qquad (5) \quad [16]$$

The length of the confidence interval was divided by 2 to obtain the δy value. Then, this value was divided by the mean percent error of calculated diffusivity in Table 2.7. The results of this analysis are shown in the following table.

# Randomly Selected Clones	δy/y	Confidence Interval Length (log-log)
10	0.16	0.32
20	0.15	0.30
30	0.14	0.28
50	0.18	0.36
100	0.12	0.24
200	0.12	0.24
300	0.10	0.20
400	0.09	0.18
500	0.08	0.16
600	0.08	0.16
1000	0.05	0.10

Table A.1. Confidence interval lengths by number of clones, fitted for log-log plot.

Using the data above, the confidence intervals were then fitted in log-log plot and the resulting plot is shown in Figure 2.6.

A.2 Worked example

For a sample of 10 randomly selected clones, the standard deviation of the accuracy of calculated diffusivity was 15.99%. It was obtained from the Matlab software. Then, the standard deviation was divided by the square root of 100 trials, since 100 different values of the accuracy of calculated diffusivity were obtained for the 10 randomly selected clones, as shown below.

$$CI_{10\ clones} = z_{95\%\ confidence}\left(\frac{15.99\%}{\sqrt{100\ trials}}\right) \quad (6)$$

Afterwards, this value was then multiplied by 1.96, which is the z-value at the 95% confidence level, as shown below.

$$CI_{10\ clones} = 1.96 \times \left(\frac{15.99\%}{\sqrt{100\ trials}}\right) \quad (7)$$

To obtain the confidence interval length, this calculated value was doubled. The result was 6.27%, as shown in Table 2.7.

For the confidence interval length in the log-log plot, the confidence interval length was divided by 2. Afterwards, Equation (5) was applied, as shown below.

$$\delta z = 0.434 \times \frac{3.14\%}{19.70\%} \quad (8)$$

Next, the δz value, which is the error value, was doubled to obtain the confidence interval length in the log-log plot, as shown in Table A.1. Here, the confidence interval length, when plotted in log-log, was 0.32 for 10 randomly selected clones.

Appendix B

Accuracy of Calculated Diffusivity by Clone Size

In this section, the accuracy of calculated diffusivity will be examined by clone size to assess its relationship. The following section details the methods used to assess this relationship.

B.1 Calculating diffusivity by clone size

From a given snapshot of the cells' clonal distribution at a fixed time point, the center of the clone of cells was located, and the distances to the center of the clone were calculated using Equation (2). Then, the mean distance to the center of the clone (DCC) was obtained, and diffusivity values were obtained from the linear model in Table 2.3 using mean DCC.

The diffusivity values were grouped by clone size using the Matlab software. The percent error of these diffusivity values was calculated by comparing them to the diffusivity values obtained from the PRW model in Equation (1), and the mean percent error was obtained for each clone size. The following table summarizes the results of this analysis.

Clone size (# of cells)	Mean % error of calculated diffusivity
4	45.83
16	37.22
32	34.61
64	33.13
128	37.01

Table B.1. Accuracy of calculated diffusivity by clone size.

The results were plotted below to analyze the relationship between accuracy of calculated diffusivity and clone size, as shown below.

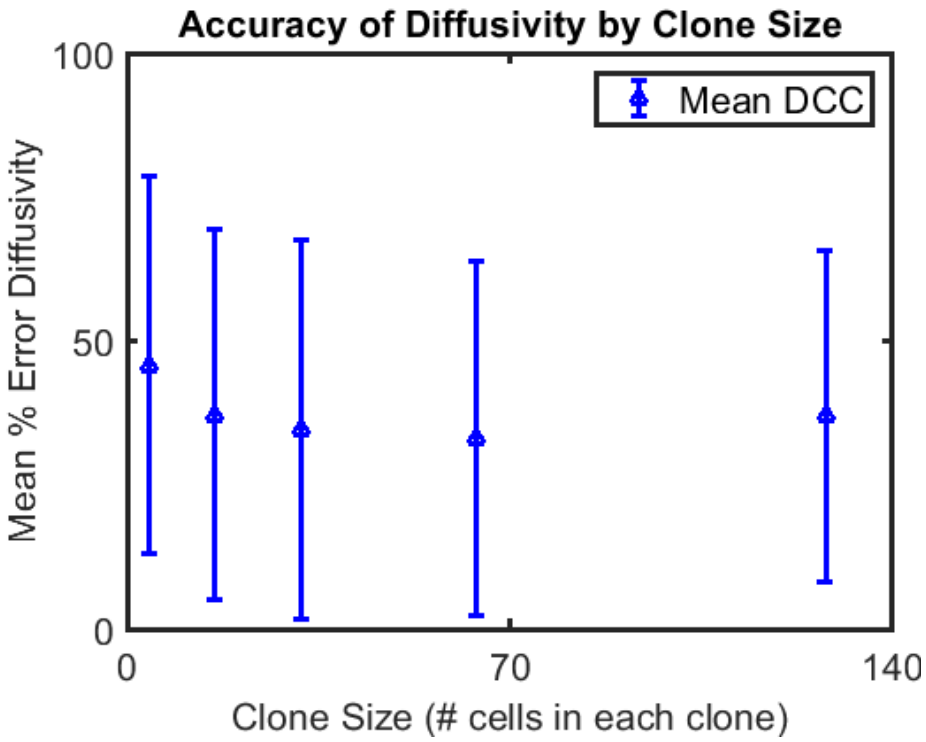

Figure B.1. Plot of accuracy of diffusivity by clone size.

The results show that accuracy of diffusivity decreased by clone size. However, there was a slight rebound from 64 clones to 128 clones. There was a large standard deviation of the accuracy of calculated diffusivity for each clone size, which could be attributed to the stochastic nature of cell migration [2].

Appendix C

Centroid of the Clone of Cells

This section aims to compare the accuracy of calculated diffusivity values obtained from the linear model using the mean and median distances to the centroid of the clone, as well as the mean square displacement from the centroid of the clone. The following sections will outline the procedures used for this analysis.

C.1 Finding the centroid of the clone

From a given snapshot of a clonal distribution of cells at a specific time point (72 hours in this case), the centroid of the clone was located by taking the mean value of all of the x and y coordinates of the cells in the clonal distribution. The mean x and y coordinates of all cells in the clone was determined to be the centroid of the clone, as shown in the figure below.

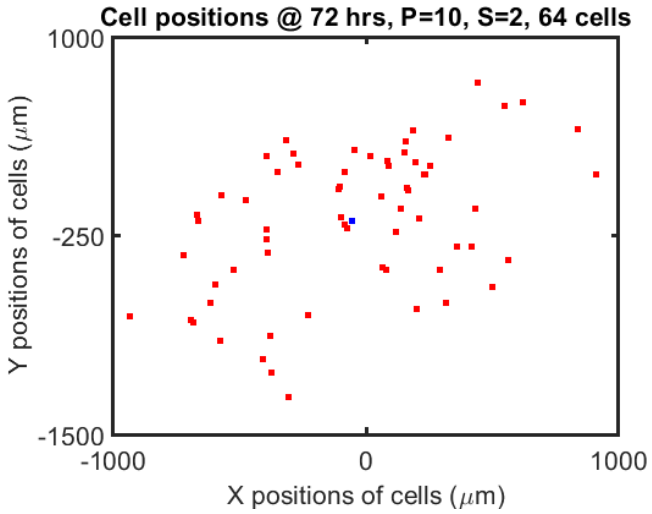

Figure C.1. Snapshot of cell positions at 72 hours with centroid of clone.

As shown, the centroid of the clone, defined as the mean x and y coordinates of all cells in the clonal distribution, is shown as a blue dot, while the cells are shown as red dots.

Using the centroid of the clone, the mean square displacement (MSD) and the distances to the centroid of the clone (DCdC) were calculated using the following equations.

$$DCC = \sqrt{(x_n - x_{centroid})^2 + (y_n - y_{centroid})^2} \quad (9)$$

$$MSD = \frac{1}{\#\ cells} \sum_{n=1}^{\#\ cells} (x_n - x_{centroid})^2 + (y_n - y_{centroid})^2 \quad (10)$$

Then, the mean and median distance to the centroid of the clone (DCdC) were obtained for each clone of cells.

C.2 Calculating diffusivity values

The MSD and the mean and median distance DCdC values for each clone of cells were fitted with the diffusivity values that were obtained from the PRW model via Equation (1), and the linear models were obtained. The fitting was done in log-log to address the skewness towards large values [8], and the linear models used were in log-log. The linear models are shown in the following table.

Observation	Linear Model Used
Mean DCdC	$\ln(\check{D})=1.88\ln(\text{mean DCdC})-8.21$
Median DCdC	$\ln(\check{D})=1.87\ln(\text{median DCdC})-8.07$
MSD	$\ln(\check{D})=0.94\ln(\text{MSD})-8.39$
MSD (theoretical)	$\ln(\check{D})=\ln(\text{MSD})-9.76$

Table C.1. Linear models used for calculating diffusivity from mean DCdC, median DCdC, and MSD from centroid.

Using the above linear models, the diffusivity values were calculated, and the mean percent error of calculated diffusivity from each method was obtained. The following table summarizes these results.

Method of Diffusivity Calculation	Mean % Error Calculated Diffusivity
Mean DCdC	38.03%
Median DCdC	40.57%
MSD	38.05%
MSD (theoretical)	41.52%

Table C.2. Accuracy of calculated diffusivity using MSD from centroid, mean DCdC, and median DCdC.

The above results show a small difference in the accuracy of calculated diffusivity when the centroid was used instead of the center of the clone.

Appendix D

Results of Calculated Diffusivity by Clone Size

This appendix provides the remaining heat maps that pertain to the results of calculated diffusivity by clone size. Note that only the clone sizes of 32 and 128 cells are used in the main text.

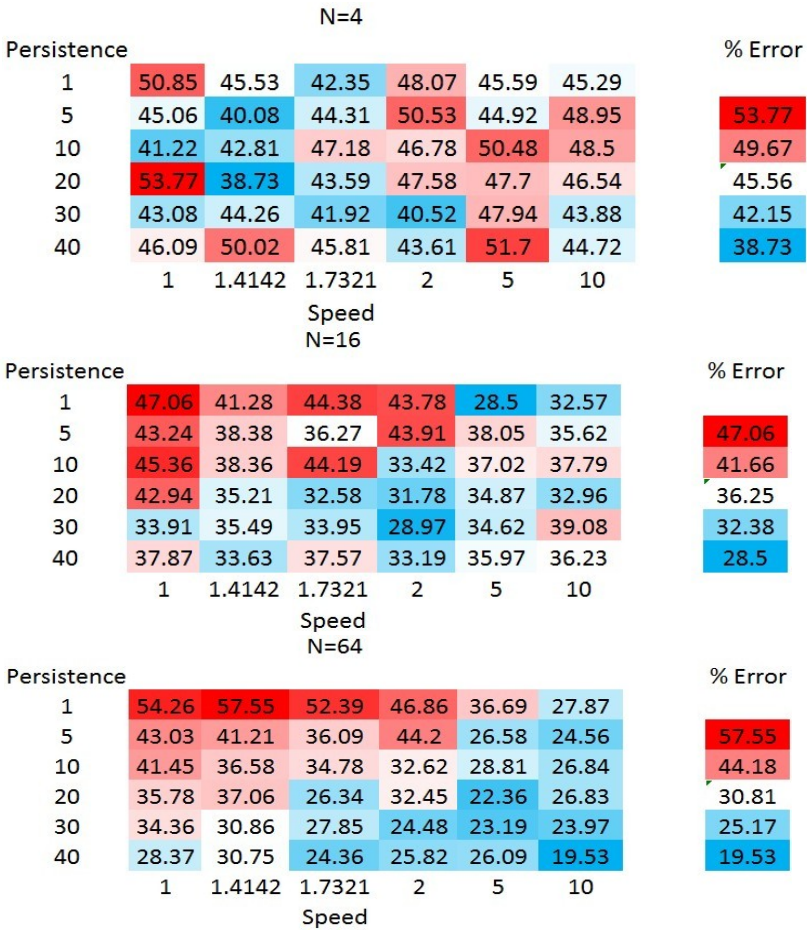

Figure D.1. Heat maps of accuracy of calculated diffusivity, obtained from clone size of 4 cells (top), 16 cells (middle), and 64 cells (bottom).

Bibliography

[1] Wirtz, D.; Konstantopoulos, K.; Searson, P.C. The physics of cancer: the role of physical interactions and mechanical forces in metastasis. *Nat. Rev. Cancer*. 2011, *11(7)*, 512-522.

[2] Wu, P.H.; Giri, A.; Sun, S.X.; Wirtz, D. Three-dimensional cell migration does not follow a random walk. *Proc. Natl. Acad. Sci. USA*. 2014, *111(11)*, 3949-54.

[3] Wu, P.H.; Giri, A.; Wirtz, D. Statistical analysis of cell migration in 3D using the anisotropic persistent random walk model. *Nat. Protoc*. 2015, *10(3)*, 517-27.

[4] Harms, B.D.; Bassi, G.M.; Horwitz, A.R.; Lauffenburger, D.A. Directional Persistence of EGF-induced Cell Migration is Associated with Stabilization of Lamellipodial Protrusions. *Biophysical Journal*. 2005, *88(2)*, 1479-1488.

[5] Staneva, R.; Barbazan, J.; Simon, A.; Vignjevic, D.M.; Krndija, D. Cell Migration in Tissues: Explant Culture and Live Imaging. In *Cell Migration*. *Methods in Molecular Biology*; Gautreau, A.; Humana Press: New York, 2018; vol. 1749, pp 163-173.

[6] Cooper, G.M. *The Cell: A Molecular Approach*, 2nd ed.; Sinauer Associates: Sunderland, 2000.

[7] pdist: Pairwise distances between pairs of observations. MathWorks. https://www.mathworks.com/help/stats/pdist.html (accessed Apr 29, 2019).

[8] When should I Use Logarithmic Scales in My Charts and Graphs? Forbes. https://www.forbes.com/sites/naomirobbins/2012/01/19/when-should-i-use-logarithmic-scales-in-my-charts-and-graphs/#71e9a7ea5e67 (accessed Apr 29, 2019).

[9] DAPI (4'-6-Diamidino-2-Phenylindole, Dihydrochloride). Thermo Fisher Scientific. https://www.thermofisher.com/order/catalog/product/D1306?SID=srch-srp-D1306 (accessed Apr 29, 2019).

[10] Gaussian Mixture Model. Brilliant. https://www.brilliant.org/wiki/gaussian-mixture-model/ (accessed Apr 29, 2019).

[11] Different covariance types for Gaussian Mixture Models. Stack Exchange. https://stats.stackexchange.com/questions/326671/different-covariance-types-for-gaussian-mixture-models (accessed Apr 29, 2019).

[12] Cluster Using Gaussian Mixture Models. MathWorks. https://www.mathworks.com/help/stats/clustering-using-gaussian-mixture-models.html (accessed Apr 29, 2019).

[13] Mohammed, E.A.; Naugler, C.; Far, B.H. Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics. In *Emerging Trends in Computational Biology, Bioinformatics, and System Biology*; Tran, Q.N., Arabnia, H.; Elsevier: Atlanta, 2015; pp 577-602.

[14] Confidence Intervals. Math is Fun. https://www.mathsisfun.com/data/confidence-interval.html. (accessed Apr 29, 2019).

[15] Confidence Intervals. Boston University Medical Center. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals_print.html. (accessed Apr 29, 2019).

[16] Estimating and Plotting Logarithmic Error Bars. University of Washington. https://faculty.washington.edu/stuve/log_error.pdf. (accessed May 1, 2019).

Curriculum Vitae

Education

2017-2019	MSE in Chemical and Biomolecular Engineering, Johns Hopkins University
2013-2017	BS magna cum laude in Chemistry, University of Pittsburgh

Awards and Activities

2017-2019	Research assistant, Wirtz Lab, Chemical and Biomolecular Engineering Department, Johns Hopkins University
2018-2019	Second Year Master's Scholarship, Chemical and Biomolecular Engineering Department, Johns Hopkins University
2016	Research assistant, Emerson Lab, Ecology and Evolutionary Biology Department, UCI
2015-2017	Research assistant, Jordan Lab, Chemistry Department, University of Pittsburgh
2015-2017	Teaching assistant, Chemistry Department, University of Pittsburgh
2015-2016	ACS Tutor, Chemistry Department, University of Pittsburgh
2014	Research assistant, Brummond Lab, Chemistry Department, University of Pittsburgh
2014-2017	Volunteer at Magee-Womens Hospital, University of Pittsburgh Medical Center
2014	Volunteer at Healthbridge Children's Hospital, Orange, CA
2013-2017	Dean's List Awards at the University of Pittsburgh
2013-2017	University of Pittsburgh Honors College Scholarship

Research Experience

2017-2019	Experiment on characterizing 2-dimensional cell migration using novel approaches with Professor Wirtz and Pei-Hsun Wu
2016	Summer research on genetic and protein composition of fruit flies with Professor Emerson
2015-2017	Computational chemistry research on molecular movement simulations with Professor Jordan
2014	Organic chemistry research with Professor Brummond

Personal

Allan B. Yue was born on March 6, 1996. He is a US citizen.

That's all!